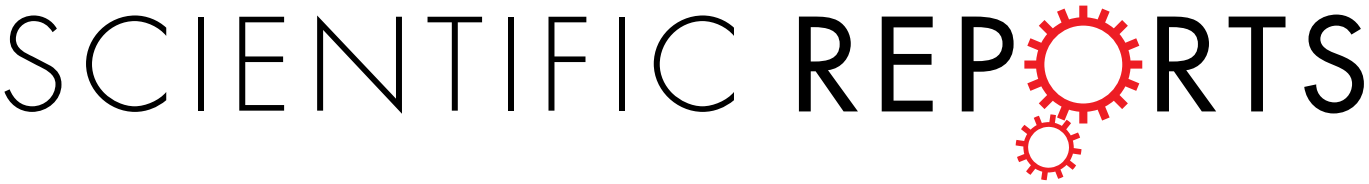
Spatiotemporal neural characterization of prediction error valence and surprise during reward learning in humans

Fouragnan, Elsa; Queirazza, Filippo; Retzler, Chris; Mullinger, Karen; Philiastides, Marios G.

DOI:
10.1038/s41598-017-04507-w



OPEN

Spatiotemporal neural characterization of prediction error valence and surprise during reward learning in humans

Received: 8 December 2016
Accepted: 17 May 2017
Published online: 06 July 2017

Elsa Fouragnan[1,2], Filippo Queirazza[1], Chris Retzler[1,3], Karen J. Mullinger[4,5] & Marios G. Philiastides[1]

Reward learning depends on accurate reward associations with potential choices. These associations can be attained with reinforcement learning mechanisms using a reward prediction error (RPE) signal (the difference between actual and expected rewards) for updating future reward expectations. Despite an extensive body of literature on the influence of RPE on learning, little has been done to investigate the potentially separate contributions of RPE valence (positive or negative) and surprise (absolute degree of deviation from expectations). Here, we coupled single-trial electroencephalography with simultaneously acquired fMRI, during a probabilistic reversal-learning task, to offer evidence of temporally overlapping but largely distinct spatial representations of RPE valence and surprise. Electrophysiological variability in RPE valence correlated with activity in regions of the human reward network promoting approach or avoidance learning. Electrophysiological variability in RPE surprise correlated primarily with activity in regions of the human attentional network controlling the speed of learning. Crucially, despite the largely separate spatial extend of these representations our EEG-informed fMRI approach uniquely revealed a linear superposition of the two RPE components in a smaller network encompassing visuo-mnemonic and reward areas. Activity in this network was further predictive of stimulus value updating indicating a comparable contribution of both signals to reward learning.

Learning occurs with the combination of two critical dimensions of decision outcome signaling discrepancies between expectations and reality: the categorical valence (i.e. whether an outcome is overall better or worse than expected) and the absolute degree of deviation from expectations ([high or low]). Traditional learning theories focusing on the appetitive domain[1,2] have treated these quantities as two facets of a single unified representation, commonly referred to as the reward prediction error (RPE) signal. Correspondingly, many animal[3,4] and human neuroimaging studies[5,6] have focused on identifying brain areas exhibiting a positive relationship with a fully parametric (signed) RPE signal. To date, however, the extent to which RPE valence and surprise (unsigned RPE)[7–9] can also be encoded separately in the human brain and the degree to which they converge (in space and time) to drive learning, remain poorly understood.

Investigating these open questions is critical as RPE valence and surprise could subserve related but largely separate functions. For instance, RPE valence determines the direction in which learning will occur. Positive RPEs will increase the likelihood of similar decisions in the future, whereas negative RPEs will increase the likelihood of an avoidance response[10–12]. In contrast, the surprise dimension modulates the amount of attention devoted to outcomes[13–15], which in turn dictates the degree of learning and ultimately the updating of future reward expectations[5,16–18]. Highly unpredicted outcomes (either positive or negative) can boost attention and

[1]Institute of Neuroscience & Psychology, University of Glasgow, Glasgow, UK. [2]Department of Experimental Psychology, University of Oxford, Oxford, UK. [3]Department of Behavioural & Social Sciences, University of Huddersfield, Huddersfield, UK. [4]Sir Peter Mansfield Magnetic Resonance Center, School of Physics and Astronomy, University of Nottingham, Nottingham, UK. [5]Birmingham University Imaging Centre, School of Psychology, University of Birmingham, Birmingham, UK. Correspondence and requests for materials should be addressed to E.F. (email: elsa.fouragnan@psy.ox.ac.uk) or M.G.P. (email: Marios.Philiastides@glasgow.ac.uk)

facilitate learning, whereas less surprising outcomes attracting reduced attention might slow down the learning process instead[19,20].

In line with these separate functional roles, recent studies in animals and humans have observed distinct neural signals aligning both with RPE valence and surprise within a distributed network of areas, including the midbrain[21,22], the striatum[18,23,24], the cingulate cortex and the anterior insula[7,9,25,26]. Specifically, these recent studies employing model-based fMRI have begun to characterize the neural basis of both signals providing new evidence that separate brain networks might encode RPE valence and surprise in the human brain[27–29]. Given the poor temporal resolution of fMRI, however, these studies preclude a rigorous assessment of the relative timing and underlying dynamics of these two learning-related signals.

Our previous electroencephalography (EEG) work revealed two distinct components encoding categorical RPE valence (early and late) and a separate outcome surprise component that overlapped temporally with the later of the two valence components[30], suggesting that the brain might require simultaneous access to both signals to drive learning. The poor spatial resolution of the EEG, however, prevented a thorough characterization of the spatial generators associated with each outcome representation. More recently, we used simultaneous EEG-fMRI, to map out the spatiotemporal dynamics of the two RPE valence components[12]. However, no analysis was presented that examined the neural systems associated with surprise or their relationship with RPE valence.

Here, using the data in ref. 12 we aim to provide a comprehensive spatiotemporal characterization of the relationship between the late temporally overlapping representations of RPE valence and surprise. Our hypothesis is that endogenous trial-to-trial variability in electrophysiologically derived measures of the two outcome variables can be used to form separate fMRI predictors to tease apart the brain networks associated with each representation. Moreover, we directly test the hypothesis that temporal and spatial congruency, of otherwise separate RPE valence and surprise representations, constitutes a viable mechanism for driving reward learning in the appetitive domain.

Results

RPE valence and surprise predicting choice behavior. To test our hypotheses and investigate the relative and potentially combined contribution (spatial and temporal) of RPE valence and surprise to learning, we present new analyses of data reported in ref. 12. Twenty participants were engaged in a probabilistic reversal-learning task[31] while we recorded simultaneous EEG-fMRI data. Specifically, on each trial subjects were presented with pairs of abstract symbols – selected from a larger set of three symbols – and through feedback (representing rewarding or non-rewarding outcomes) learned to select the one with the highest reward probability. At any given point during the experiment, one of the three symbols was associated with a 70% chance of obtaining a reward ("high" reward probability symbol) compared to a 30% chance for the remaining two symbols ("low" reward probability symbols). After reaching a predefined learning criterion, the high reward probability was re-assigned to a different symbol and subjects had to enter a new learning phase (see Fig. 1a and Methods). Participants' behavior was probabilistic and mirrored the principles of a RL mechanism[1] (Fig. 1b) best explained with a simple model-free RL mechanism with a fixed learning rate (see model comparisons in the Methods section).

To investigate whether RPE valence and surprise are better predictors of behavior than the signed RPE itself, we first modeled behavior by considering i) stay/switch choice patterns and ii) reaction time (RT) changes (i.e. Delta RTs; denoting either acceleration or deceleration) on subsequent trials, using mixed-effects logistic and linear regression analyses respectively (see Methods). Correspondingly, we performed formal model comparisons using a likelihood ratio test and found that a model including an interaction between RPE valence and surprise explained behavior better than a model including only the signed RPE (Stay/switch behavior: $\chi^2_{(2)} = 36.24$, $P < 0.001$; Delta RTs: $\chi^2_{(2)} = 18.81$, $P < 0.001$). This result is in line with the growing consensus that separate learning systems are at play in the human brain[28].

More specifically, we found that participants were more likely to repeat the same choice after positive compared to negative feedback ($\chi^2_{(1)} = 35.67$, $P < 0.001$). Moreover, a significant interaction between RPE valence and surprise ($\chi^2_{(1)} = 21.5$, $P < 0.001$) (see Fig. 1c) revealed that the probability of repeating the same choice in the next trial was negatively related to surprising *positive RPE*s and positively related to surprising *negative RPE*s, replicating findings from previous studies using the same experimental design[30]. Due to our design, highly surprising *positive RPE*s are generated mainly following positive feedback in trials with pairs of low reward probability symbols and thus participants are – on average – more likely to switch back to the high reward probability symbol on subsequent trials. In contrast, highly surprising *negative RPE*s come about primarily following negative feedback on trials where the high reward probability stimulus was chosen and thus participants are more likely – on average – to continue choosing the high reward probability symbol on subsequent trials.

Moreover, we found that positive RPEs speed up subsequent approach behavior whereas negative feedback slows down subsequent response times (main effect of valence on Delta RTs: $F_{(1, 28.04)} = 5.9$; $P = 0.021$). Notably, there was additionally a significant interaction between RPE valence and surprise ($F_{(1, 17.89)} = 33.45$; $P < 0.001$, Fig. 1d) such that whilst higher surprising *positive RPE*s resulted in faster subsequent responses, higher surprising *negative RPE*s resulted in slower response times. Very intuitively, this result implies that, after participants have experienced an unexpected negative RPE, they strategically slow down on the subsequent trial to make more optimal choices but at the expense of longer RTs.

Temporal superposition of RPE valence and surprise. To identify temporal components encoding RPE valence and surprise we used a multivariate discriminant analysis of feedback-locked EEG signals[32–34]. Specifically, for each participant, we estimated linear weightings of the EEG electrode signals that maximally discriminated between conditions of interest (see below) over several distinct temporal windows (Eq. 4). Applying the estimated electrode weights to single-trial data produced a measurement of the discriminating component

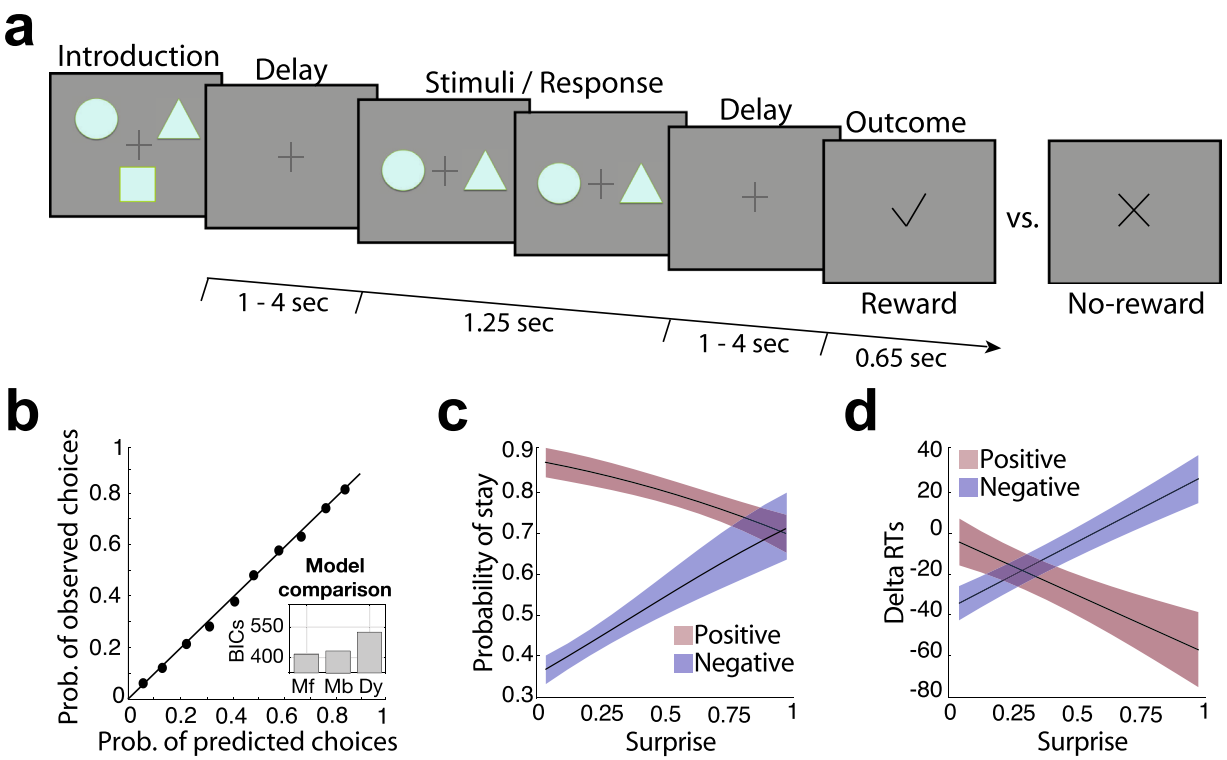


Figure 1. Schematic representation of the experimental task and modeling of behavioral responses. (**a**) Each trial began with a random delay followed by the presentation of two abstract symbols (selected from a larger set of three symbols) for a period of 1.25 s. During this time, subjects pressed one of two buttons on a response device to indicate which of the two symbols (right or left) they believed was more likely to lead to a reward. The fixation cross flickered for 100 ms when a selection was made. Finally the decision outcome was revealed after a second random delay. A tick or a cross were used to inform the participants of a positive or a negative outcome, respectively. (**b**) Main panel: Model-predicted choice probabilities (x-axis) derived from a RL algorithm using a softmax procedure (binned into ten bins – bin size of 0.1 - and averaged across all subjects and across symbols) closely matched participants observed behavioral choices (y-axis), calculated for each bin as the fraction of trials in which they chose one of the three symbols. Small panel: A Bayesian model comparison using BIC scores revealed that the Model-free (Mf) RL model explained the data better than a Model-based (Mb) RL model and a model including a dynamic learning rate (Dy) (Lower BIC scores indicate a better fit – see Methods). (**c**) A mixed-effects regression analysis demonstrated that participants were more likely to repeat the same choice after positive compared to negative feedback with a significant interaction between RPE valence and surprise (see text for details). (**d**) A mixed-effects regression analysis revealed that positive and negative RPEs speed up and slow down reaction times on subsequent trials (Delta RTs = RTs(t + 1) − RTs(t)) respectively, with a significant interaction between RPE valence and surprise (see text for details).

amplitudes (henceforth y), which represent the distance of individual trials from the discriminating hyperplane (Fig. 2a). We subsequently used the single-trial variability (STV) in these values to construct parametric fMRI regressors to identify the brain regions that correlate with the resultant discriminating EEG components (Fig. 2b). To quantify the discriminator's performance over time we used the area under a receiver operating characteristic curve (i.e. A_z value) with a leave-one-out trial cross validation approach.

Using this approach, we previously established the presence of two temporally specific EEG components discriminating – reliably in individual participants – between positive and negative RPEs in the time range 180–370 ms after an outcome (Fig. 3a)[12]. The second of those valence components peaked on average 308 ms (SD ± 37.7) after outcome and was the only component that was directly linked to reward learning by either up- or down-regulating the human reward network to update the expected value of the stimuli following positive and negative RPEs respectively (the first valence component reflected an early alertness response and was selective for negative RPEs only). Importantly, this Late RPE valence component was decoupled (i.e. uncorrelated) from outcome surprise replicating our previous work[12] (mixed-effects regression on outcome surprise: interaction effect between positive/negative RPEs and STV in Late valence component: $F_{(1, 5.65)} = 0.01$; $P = 0.89$; Pearson's correlation coefficients: Negative RPE: $r = 0.01$; Positive RPE: $r = 0.03$ and see Fig. 3b). This result highlights that our Late RPE valence EEG component was not additionally modulated by surprise, suggesting that a fully signed RPE representation in the brain is largely subcortical in nature and therefore not contributing strongly to our EEG signal.

In this work, we ran a separate multivariate analysis of the EEG to directly discriminate along an outcome surprise dimension, by discriminating between trials with very high (in the range [0.8–1]) and very low (in the

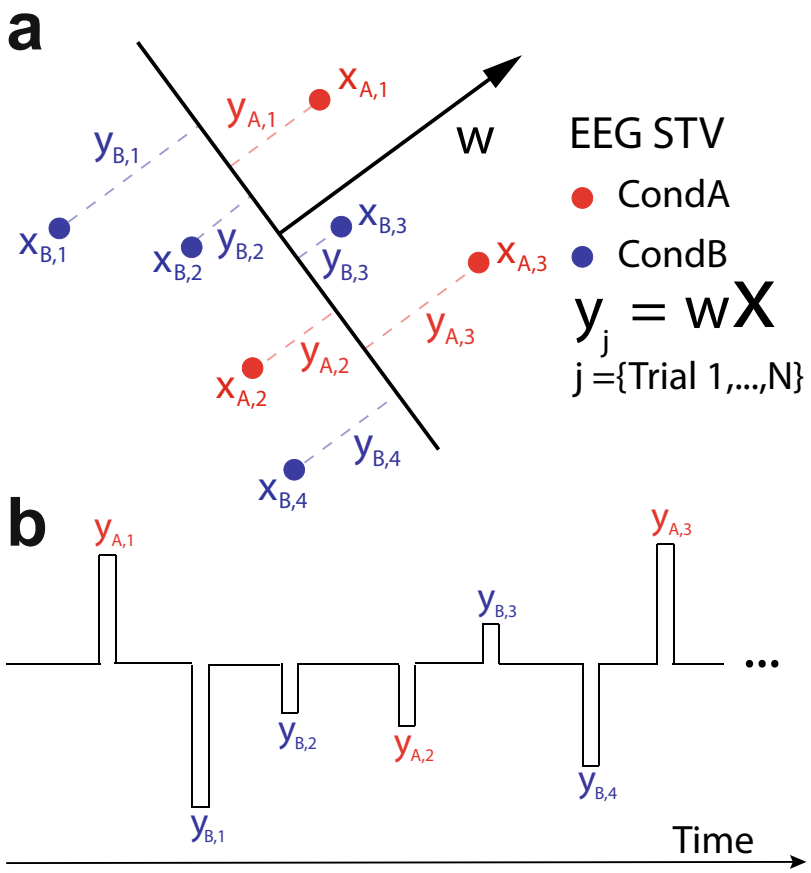


Figure 2. Single-trial discriminant analysis and EEG-informed fMRI regressors. (**a**) We used single-trial analysis of the EEG to perform binary discriminations between conditions of interest, here denoted as condition A and B (in red and blue respectively). We first estimated *w*, which is a linear weighting on the EEG sensor data (*X*) that maximally discriminates between the two conditions. This determines a task-related projection (*y*) of the data, in which the distance to the decision boundary reflects the decision certainty of the classifier in separating each of the relevant conditions. We treated the single-trial *y* amplitudes (single-trial variability [EEG STV]), as an index of how each condition of interest was encoded on individual trials. (**b**) Given these *y* values and their corresponding outcome-locked onset time points, we built fMRI regressors for subsequent GLM analyses. These regressors were all convolved with the canonical HRF. Details of specific events included in each EEG-informed fMRI regressor can be found in the main text (see *fMRI analysis* section).

range [0–0.2]) unsigned RPE. In doing so, we identified a separate EEG component that peaked, on average, at 320 ms (SD ± 32.1) after outcome (Fig. 3c). The timing of this component overlapped with that of the second RPE valence component presented above (see shaded regions in Fig. 3a and c). We formally tested if the peak times of the two EEG components were significantly different across individuals and found no significant differences (paired t-test, $t_{19} = 1.05$, $P = 0.31$), suggesting that the brain might require near simultaneous access to both signals to drive learning.

To test whether this additional EEG component was indeed parametrically modulated by outcome surprise (rather than responding categorically to high vs. low surprise – unsigned RPE), we computed discriminator amplitudes for trials with intermediate surprise levels (i.e. high unsigned RPE [0.6 – 0.8]; medium unsigned RPE [0.4–0.6] and low unsigned RPE [0.2–0.4]), which were not originally used to train the classifier ("unseen" data). Specifically, we applied the spatial weights of the window that resulted in the highest discrimination performance for the extreme outcome surprise levels to the EEG data with intermediate values. We expected that these "unseen" trials would show a parametric response profile as a function of outcome surprise and therefore the resulting mean component amplitude at the time of peak discrimination would proceed from very low < low < medium < high < very high surprise.

Correspondingly, we found that the mean discriminator output (*y*) for each category appeared to parametrically increase as a function of the unsigned RPE signal (Fig. 3d, Grey: Intermediate categories, Yellow: Categories used for discrimination). We further confirmed the linear relationship between the STV in the EEG RPE surprise component and its model-based counterpart using a single-trial mixed-effects regression analysis ($F_{(1,\ 17.08)} = 64.57$; $P < 0.001$). Though this relationship was significant, the single-trial correlation between our EEG component amplitudes and the model-derived estimates of unsigned RPE was moderate (average Pearson's correlation coefficient across subjects: $r = 0.31$). This finding suggests that the STV in our electrophysiologically-derived measure of surprise could

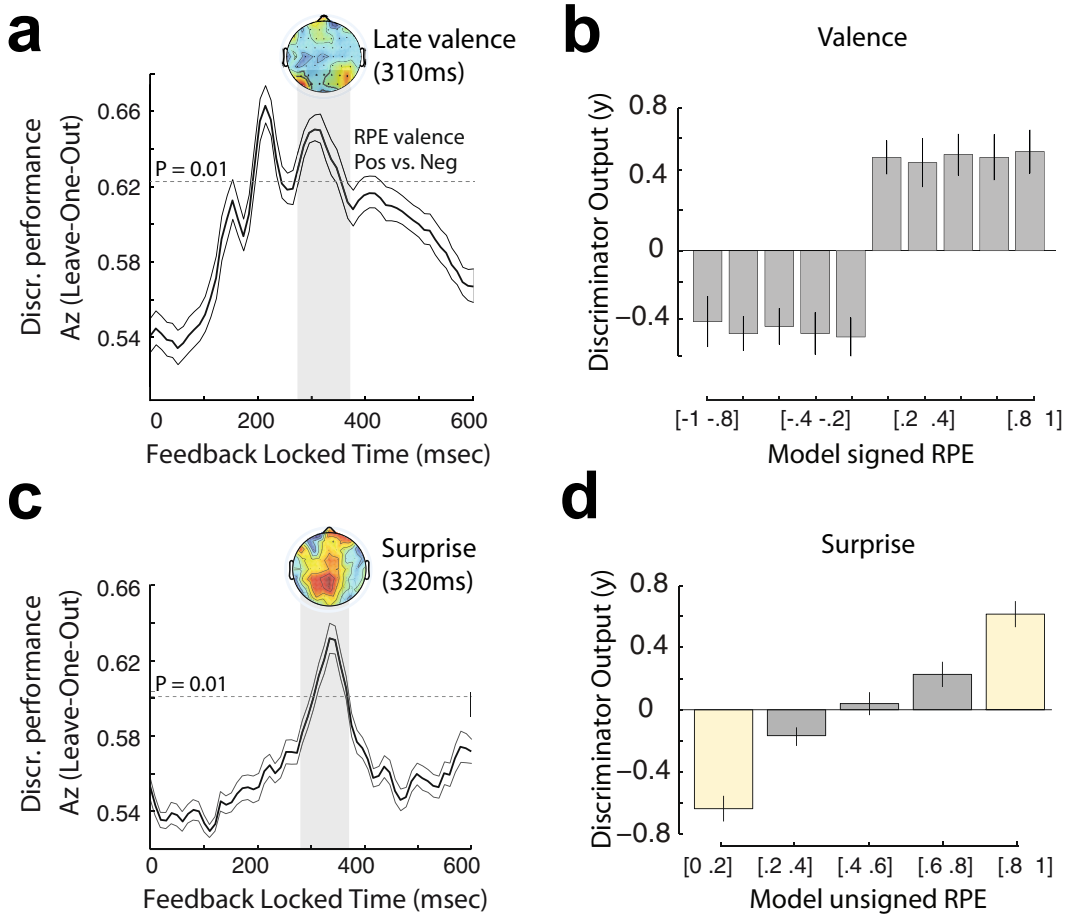


Figure 3. Single-trial EEG analyses. (**a**) Discriminator performance (cross-validated A_z) during RPE valence discrimination (positive vs. negative feedback) of outcome-locked EEG responses, averaged across subjects ($N = 20$). The dotted line represents the average A_z value leading to a significance level of $P = 0.01$, estimated using a bootstrap test. Shaded error bars are standard errors across subjects. In this work, we are focusing only on the second of the two RPE valence components (Late Valence component). The scalp map represents the spatial topography of this component. (**b**) Mean discriminator output (y) for the Late valence component, binned in ten quantiles based on model-based signed RPE estimates. This component exhibits mainly a categorical response profile without an additional modulation by the magnitude of the RPE. (**c**) Discriminator performance (cross-validated A_z) during RPE surprise discrimination (very low vs. very high surprising outcomes), averaged across subjects ($N = 20$). The dotted line represents the average A_z value leading to a significance level of $P = 0.01$, estimated using a bootstrap test. Shaded error bars are standard errors across subjects. The scalp map represents the spatial topography of the outcome surprise component. **(d)** Mean discriminator output (y) for the outcome surprise component, binned in five quantiles based on model-based unsigned RPE estimates, showing a parametric response along the outcome surprise dimension. Yellow bins indicate trials used to train the classifier, while grey bins contain "unseen" data with intermediate outcome surprise levels. Error bars are standard errors across subjects.

offer additional explanatory power in our subsequent fMRI analysis (i.e. in addition to its model-derived counterpart) to enable a more comprehensive characterization of the brain networks involved in encoding outcome surprise.

To reinforce the notion that the EEG STV in this component carried task-relevant information, which we could exploit further in the fMRI analysis, we performed an additional single-trial mixed-effects regression analysis. Specifically, we showed that the trial-by-trial fluctuations in our outcome surprise component were predictive of the single-trial learning rates in a version of our RL model with a dynamically updating learning rate parameter[31,35–37] (see Methods). In other words, the higher the surprise (as indexed by the EEG), the more recent observations were considered for updating the stimulus value for the next decision, thereby boosting the speed of learning and thus increase the single-trial learning rates ($F_{(1,\ 11.99)} = 13.95$; $P = 0.002$). Ultimately, this result offers a more concrete link between behavior and neuronal variability in our EEG outcome surprise component.

Spatial signatures of RPE valence and surprise. Our main fMRI analysis was designed to expose the neural correlates of the temporally overlapping representations of RPE valence and surprise in a whole-brain

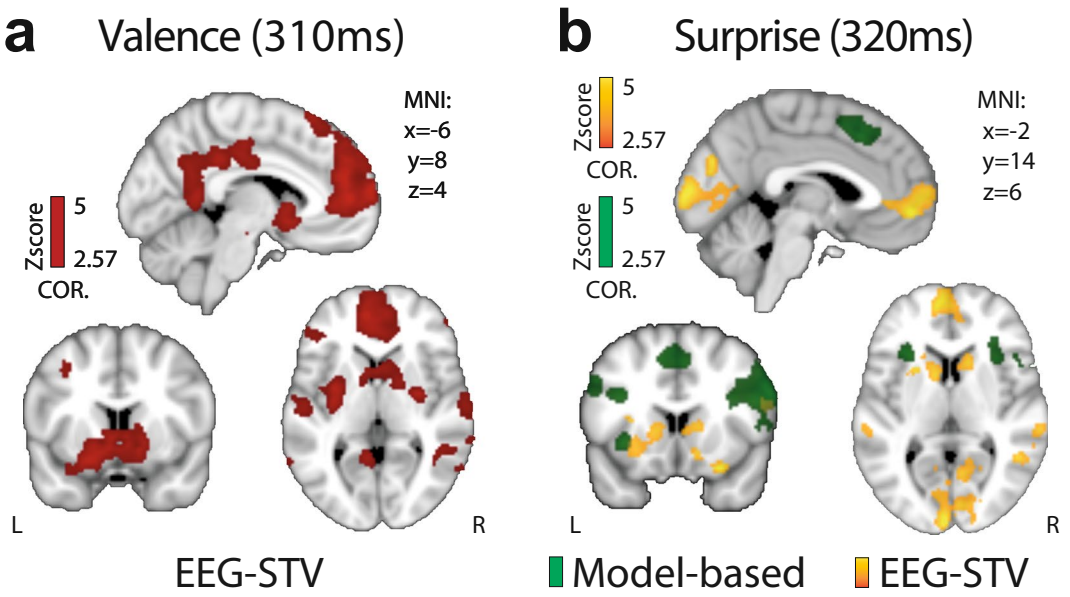


Figure 4. Spatiotemporal characterization of RPE valence and surprise signals. (**a**) Regions correlating with the EEG STV in our valence component, exhibiting overall greater response for positive compared to negative RPEs. (**b**) Regions correlating positively with outcome surprise as captured by a RL model (green) and the STV in our corresponding EEG component (yellow), respectively. Note the complementary nature of activations in the EEG STV map. All activations represent mixed-effects and are rendered on the standard MNI brain at $Z > 2.57$, cluster-corrected ($P < 0.05$) using a resampling procedure (minimum cluster size = 76 voxels).

analysis by using the explanatory power of the STV in the EEG components associated with these representations. We hypothesize that endogenous variability in these components in individual participants can carry additional information about the internal processing of RPE valence and surprise usually estimated with a simple categorical contrast (positive-vs-negative RPEs) and a model-based unsigned RPE regressor, respectively.

Despite the temporal superposition of the RPE valence and surprise components, their EEG scalp topographies looked qualitatively different (Fig. 3a and c), indicating that the two signals might be encoded together in time but in largely different networks. In addition, as a result of the aforementioned analysis, trial-by-trial amplitude variations in the two discriminating components were largely uncorrelated, allowing us to use the endogenous STV in the component amplitudes to build parametric, EEG-informed, fMRI regressors to identify the brain networks correlating with each component. Thus, our main GLM included parametric regressors for RPE valence using the STV in the relevant EEG component and parametric regressors for outcome surprise using the model-based unsigned RPE estimates and the STV in the relevant EEG component (see Methods for a full description of the GLM).

Our EEG-informed fMRI analysis revealed largely separate and distributed networks encoding RPE valence and surprise information. As we demonstrated recently[12], the endogenous variability in the valence component covaried with a number of areas of the human reward network[38–40], such as the ventromedial prefrontal cortex (vmPFC), the striatum (STR), the amygdala, the dorsal posterior cingulate cortex (PCC), the ventral PCC, the putamen, the lateral orbitofrontal cortex (lOFC) and the posterior insula (INS) (Left: −40, −2, 8) (Fig. 4a and Table 1 for whole-brain results). In addition to these known reward structures, we also identified significant clusters in the lingual gyrus (LG), the middle temporal gyrus (MTG) and the precuneus, areas that have previously been implicated in memory retrieval and adaptive visual learning processes[39]. Activity in this rather distributed network was overall found to be both suppressed and activated in response to negative and positive RPEs respectively, an activity pattern consistent with a role in motivating both avoidance and approach learning[12,41].

To decipher how the encoding of outcome surprise is represented in the brain relative to the encoding of RPE valence, we used separate parametric fMRI predictors for surprise that were derived from a RL model (i.e. estimated purely from behavior) and EEG STV (i.e. endogenous variability) in our surprise component, respectively. Our conventional model-based fMRI regressor correlated significantly with activity in the dorsolateral PFC (dlPFC), the bilateral anterior INS (aINS), the medial PFC (mPFC), the inferior frontal gyrus, the supramarginal gyrus, the precentral gyrus and the angular gyrus (Fig. 4b; yellow clusters and Table 1 for the whole-brain results), consistent with previous reports[27]. In particular, activity in the middle frontal gyrus and insular cortex have been consistently linked to deviations from expectations in a large range of learning tasks[42,43].

Crucially, using our EEG-informed surprise regressor revealed additional activations over and above what was already conferred by its model-based counterpart (paired t-tests, all $P < 0.05$). These activations included the dorsal STR, the vmPFC, the lOFC, the LG, the bilateral MTG, the occipital pole and the precuneus (Fig. 4b; green clusters and Table 1 for the whole-brain results). As our two regressors for surprise were partially correlated (Fig. 3d) we also repeated this analysis by orthogonalizing our EEG STV regressor with respect to the model-based one (such that any common variance was absorbed by the latter) and obtained similar results. We

Region	Hemisphere	BA	Peak MNI coordinates (mm)			Z Value
			x	*y*	*z*	
Main GLM						
Late Valence STV-EEG (corrected P < 0.05)						
Ventromedial prefrontal cortex	L	10	−6	52	6	3.96
Striatum (Nucleus Accumbens)	R	—	8	10	−6	3.55
	L	—	−8	10	−8	3.58
Amygdala	R	28/36	26	−2	−22	2.71
	L	28/34	−20	−4	−18	3.63
Dorsal posterior cingulate cortex	L	23	−8	−38	32	3.54
Ventral posterior cingulate cortex	L	29	−2	−50	20	3.27
Anteromedial prefrontal cortex	L	32	−8	50	30	3.51
Superior medial prefrontal cortex	L	8	−12	28	60	3.48
Postcentral gyrus	R	43	62	−12	28	3.32
	L	43	−56	−12	28	3.48
Putamen	R	—	−30	−6	2	3.66
	L	—	−28	−6	2	3.52
Posterior Insula	L	48	−40	−2	8	2.98
Lateral orbitofrontal cortex	R	38	4	34	16	3.2
	L	38/48	−24	12	−20	2.81
Precuneus	L	—	−10	−56	30	3.56
	L	30	−10	−56	30	3.13
Middle temporal gyrus	R	20/37	56	−46	−8	2.85
	L	37	−64	−48	−6	2.78
Angular gyrus	R	39	44	−66	34	4.07
	L	39	−50	−64	40	2.93
Late Surprise STV-EEG (corrected P < 0.05)						
Dorsal striatum	R	—	12	16	6	3.68
	L	—	−10	12	6	3.47
Ventromedial prefrontal cortex	L	10	−6	54	0	4.05
Lateral orbitofrontal cortex	R	47	28	30	−14	3.81
	L	46	−34	30	14	3.51
Lingual gyrus	R	17	12	−62	12	3.36
Middle temporal gyrus	R	22	60	−32	2	3.46
	L	21	−64	−30	2	3.24
Occipital pole	R	18	6	−86	12	3.19
	L	19	−24	−90	16	2.87
Precuneus	L	40	−38	−44	42	3.61
Model-based Unsigned RPE (corrected P < 0.05)						
Middle frontal gyrus	R	44	48	12	34	3.25
	L	44	−50	8	38	3.49
Anterior Insula	R	48	34	18	0	3.13
	L	48	−36	20	0	4.02
Supplementary motor area	R	8	2	20	52	3.76
Middle temporal gyrus	R	21	62	−26	−6	3.28
Inferior frontal gyrus	R	44	52	10	18	3.69
Inferior temporal gyrus	R	37	−40	−60	−10	3.73
Supramarginal gyrus	R	40	40	−38	38	3.41
	L	40	−40	−48	42	3.46
Precentral gyrus	R	6/44	38	4	34	3.41
	L	6	−52	0	34	3.61
Angular gyrus	R	40	40	−48	40	3.48
	L	40	−46	−54	42	3.41

Table 1. Model-based and EEG-informed fMRI analyses. Complete list of activations correlating with the single-trial variability in the model-based unsigned RPE and the EEG-derived Late Valence and Surprise regressors (GLM 1; mixed effects, $|Z| > 2.57$, cluster-corrected for multiple comparisons). MNI, Montreal Neurological Institute; L, left hemisphere; R, right hemisphere, BA, Broadmann Area.

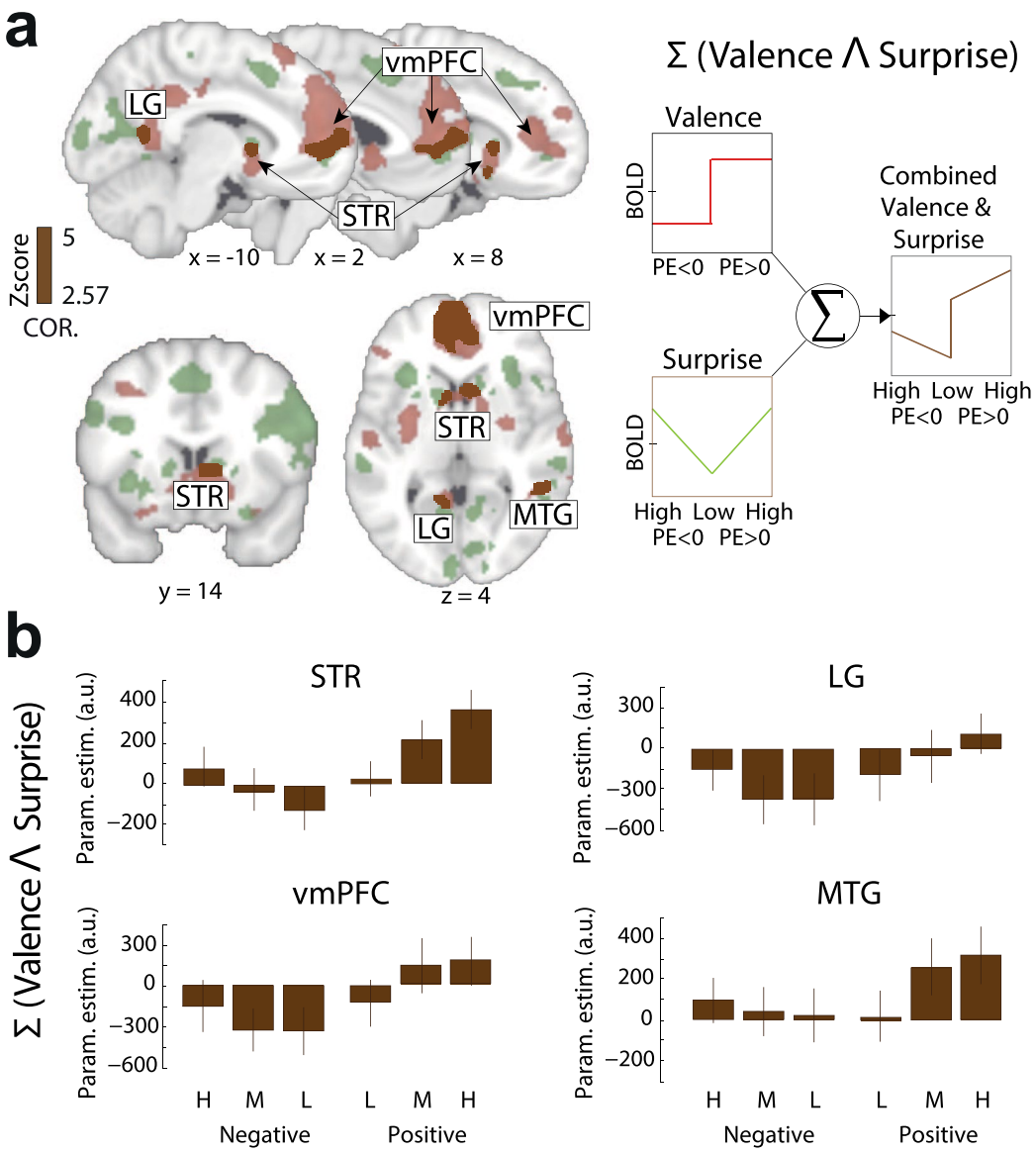


Figure 5. Full spatial representation of the RPE valence and surprise networks and their overlap. (**a**) A conjunction analysis on the results arising from the EEG-informed regressors for the RPE valence (red clusters) and surprise (green clusters) components revealed that four areas – the STR, vmPFC, LG and MTG – significantly encoded both quantities (brown clusters). The conjunction analysis was performed using the resulting whole brain activation maps for RPE valence and surprise and applying a $Z > 2.57$, cluster-corrected ($P < 0.05$) using a resampling procedure. (**b**) The four overlapping regions exhibited a clear linear superposition profile between the two RPE components with a higher BOLD signal for positive vs. negative RPEs but also a systematic increase from low (L), to medium (M), to high (H) outcome surprise trials, within each outcome type.

view this as further evidence that the EEG STV can be used to reveal relevant brain regions and that the approach can be used to complement model-based fMRI.

Temporal and spatial congruency between RPE valence and surprise. Our temporally overlapping representations of RPE valence and surprise suggest that the brain might require near simultaneous access to both signals to drive learning. This notion is also in line with classical RL theory which posits that reward learning is driven by a combination of both categorical valence and unsigned RPE information[44]. We thus, tested whether, within the largely separate spatial representations of RPE valence and surprise (Fig. 4a and b), there was a subset of regions in which both outcome dimensions were linearly combined.

To this end, we ran a conjunction analysis across statistical maps identified in our previous EEG-informed GLM. Specifically, among the unique areas found in the outcome surprise network (covarying with the EEG STV), four regions showed a significant overlap with the valence network, namely vmPFC, STR, MTG and LG (see Fig. 5a for the full spatial representation of both networks). While converging evidence emphasizes the role of

the STR and the vmPFC in reward learning[36], this result provides new evidence that additional areas such as the MTG and LG are implicated in processing reward learning. Noteworthy is that all four of these regions appeared in the conjunction analysis with our EEG-informed rather than the model-based predictor of unsigned RPE, further highlighting the utility of our simultaneous EEG-fMRI measurements in revealing latent brain states and complementing traditional model-based fMRI approaches. To further validate this point, we ran a separate GLM including only the model-based unsigned RPE and a conventional RPE valence regressor ($+1$ and -1 for positive and negative RPEs respectively) and conducted a similar conjunction analysis. We found that none of the four regions we identified above appeared in the conjunction analysis of the conventional RPE valence and unsigned RPE regressors (all P values > 0.5).

To visualize the signal pattern in these areas, we carried a separate ROI analysis and binned the trials in six categories based on RPE valence (positive or negative RPEs) and surprise (low, medium or high based on our EEG STV of the corresponding component). Note that this analysis was performed for illustrative purposes only (the ROIs were formally linked to RPE valence and surprise in the previous EEG-informed fMRI analysis). We found that activity in these regions exhibited a profile consistent with a linear superposition of the two outcome dimensions with the fMRI signal being overall higher for positive vs. negative RPEs (see illustration in Fig. 5a), while within each of the two outcome types was further modulated by surprise (Fig. 5b). This linear superposition is noteworthy because it points to a common network for the expression of both RPE valence and surprise. This interpretation is consistent with previous evidence that separate families of intermixed neurons are expressed within the same brain regions in response to both RPE valence and unsigned RPE[21].

To offer additional evidence linking these areas with reward learning, we performed an additional analysis. Specifically, we hypothesized that activity in a network reflecting both RPE valence and surprise should drive learning by encoding the value update required for the chosen option. To test for this formally we used the percent signal change (PSC; Eq. 8) in each of the four brain areas as a predictor of expected value updates estimated with our RL model in a linear regression model (Eq. 9). This analysis revealed that all regions were significantly predictive of the update in expected value at time of outcome (vmPFC: $P < 0.01$, $t_{19=}2.79$, STR: $P < 0.05$, $t_{19} = 2.06$, LG: $P < 0.001$, $t_{19} = 3.78$, MTG: $P < 0.03$, $t_{19} = 2.29$).

Discussion

Here we provided a comprehensive spatiotemporal characterization of the neural correlates of RPE valence and surprise in the human brain by capitalizing on the high temporal and high spatial resolution of simultaneously acquired EEG and fMRI measurements. More specifically, by identifying temporally-specific EEG components of RPE valence and surprise and using the trial-by-trial amplitude fluctuations in these components as predictors in an fMRI analysis, we were able to demonstrate temporally overlapping but largely spatially separate representations for the two outcome dimensions. Thus by assigning temporal order to spatial networks, our work adds to the large body of related fMRI work[7,9,27,45] by offering a more complete *spatiotemporal* representation of the relevant networks.

As we demonstrated recently[12], electrophysiological variability in an RPE valence component was found to differentially suppress or activate regions of the human reward network in response to negative and positive RPEs respectively, consistent with a role in guiding approach and avoidance learning respectively[12,41,46]. To examine the spatial extend of outcome surprise relative to this RPE valence representation, we used separate parametric fMRI predictors for unsigned RPE that were derived from a RL model and from the endogenous STV in a relevant EEG component. Our conventional model-based fMRI predictor revealed a distributed network of activations including the dlPFC, aINS and mPFC, in line with previous reports in the literature[19,27,37,47]. However, recent related findings claim that activity in regions other than those found in the conventional unsigned RPE network such as the STR or the vmPFC, might also encode surprise information at time of outcome[9,37,48,49]. Similarly, others have endorsed the notion of a ventral-dorsal gradient in the striatum encoding valence-surprise RPE information[27,50].

Consistent with these recent claims, our EEG-informed outcome surprise regressor exposed additional unique areas, including the dorsal STR, the vmPFC and the lOFC among others. This finding endorses the hypothesis that trial-by-trial variability in our electrophysiologically-derived measure of outcome surprise (i.e. endogenous variability) may carry additional explanatory power compared to its behaviorally derived counterpart (i.e. external variability). It also highlights the utility of the simultaneous EEG-fMRI measurements in exposing latent brain states, thereby complementing more conventional model-based fMRI analyses.

Overall, these new results imply that surprise is encoded near simultaneously with RPE valence but in largely separate systems, reinforcing the notion that these outcome dimensions subserve related but separate functions. More specifically, our results are consistent with the idea that the late system encoding RPE valence determines the direction in which learning occurs (approach or avoidance learning), as it is up- and down-regulated following positive and negative RPEs respectively. In contrast, the system encoding outcome surprise captures the absolute discrepancy in stimulus-reward associations, thus being responsible for speeding up or slowing down the learning process[37]. In fact, the outcome surprise network included many regions associated with the human attentional network, consistent with an increase in resource allocation for unexpected outcomes in order to facilitate learning[27,37,48].

Despite the seemingly separate spatial representations associated with the RPE valence and surprise components we also found a spatial overlap in a smaller network comprising the STR, vmPFC, MTG and LG, with activity in these regions being predictive of stimulus value updating. Importantly, this overlap appeared exclusively in the conjunction of activations appearing in the EEG-derived measures of valence and surprise. While a large body of evidence have implicated the STR and vmPFC in reward learning[36,51] previous work in primates and humans offered contradicting evidence on whether these regions encode valence *or* unsigned RPE[52,53]. Our results are intriguing as they reconcile previous reports and provide compelling evidence implicating the STR and vmPFC in processing both RPE valence and surprise.

In addition, our results provide further evidence that additional areas such as the MTG and LG are implicated in processing reward-based learning. This finding is consistent with some studies offering converging evidence that activity in MTG and LG signals behavioral switches in choice in non-human primates[54,55] and in humans[39,51]. However, another interpretation would be that activity in these visuo-mnemonic areas facilitate effective inference in changing environments[11,56] by modulating the stimulus-reward association established in memory[51]. Other studies have also reported an interplay between the vmPFC and MTG[57,58] such that RPE information can be attributed to the visual characteristics of the stimuli under consideration[51,56]. Thus existing visual memories of the stimuli could be reactivated at time of outcome and updated with new rewarding information in order to help the decision maker select the optimal alternative in subsequent trials[53,59]. Confirming this hypothesis, in a supplementary analysis, we found that the same four areas were parametrically modulated by the value of the chosen option in the decision phase, in line with previous reports[46,60,61]. Here, our results confirm the role of these two visuo-mnemonic regions in encoding stimulus value updating during learning and thus facilitating adaptive decisions.

Moreover, the fMRI signal in all four regions (STR, vmPFC, MTG and LG) exhibited a linear superposition profile of the two outcome-related signals. Although this profile is consistent with recent theories postulating a single population of neurons encoding an integrated representation of both signals in these regions[62], our finding does not preclude the possibility of separate but *intermixed* populations of RPE valence and surprise coding neurons in this network. Although speculative, this hypothesis would predict that the activity of different populations of neurons randomly distributed within an area is averaged out within fMRI voxels to give rise to the response profile we observed here. In turn this would suggest that RPE valence and surprise could continue to make independent contributions to learning. Recent evidence from electrophysiological investigations in animals supports this idea and demonstrates that valence and surprise signals are encoded by an equivalent amount of intermixed DA neurons with orthogonal coding patterns in the midbrain[21,63–65] and in the posterior parietal cortex, including the MTG[21,22].

Similarly, this result is noteworthy because it provides the first evidence in humans, to our knowledge, that a composite signal – that does not rely entirely on a single RPE representation encoded by one family of dopamine (DA) neurons[66–69] – is necessary to drive learning. This interpretation is further supported by recent evidence failing to uncover a true RPE in the nucleus accumbens, especially when categorical RPE valence is accounted for[70], highlighting the collinearity confound between the two signals in conventional fMRI designs. Although our study cannot rule out that a fully parametric signed RPE exists subcortically – as these representations are unlikely to contribute strongly to the EEG signal – our results are more consistent with another recent study reporting multiple outcome-related signals converging onto the STR[71].

In summary, our data advance our understanding of the neurobiological mechanisms of learning associations between stimuli and outcomes and suggest complementary roles for RPE valence and surprise in decision making that can help constrain formal learning theories.

Methods

Participants. Twenty-four participants took part in the study, which was conducted in accordance with approved guidelines. Informed consent was obtained from all participants while all experimental protocols were approved by the School of Psychology Ethics Committee at the University of Nottingham. Four participants were excluded due to excessive head movements inside the scanner. The remaining twenty participants (12 females; average ± SD age, 21 years ± 2.6 years) were used in all subsequent analyses. All were right handed, had normal or corrected-to-normal vision and reported no previous history of neurological problems.

Stimuli. Our task employed twelve abstract stimuli (examples given in Fig. 1a) and two feedback symbols (a tick and a cross for positive and negative feedback respectively). All stimuli (180 × 180 pixels) including those used for the feedback (125 × 125 pixels) and the fixation cross (30 × 30 pixels) were equated for luminance and contrast. The task was programmed with the Presentation software (Neurobehavioral Systems Inc., Albany, CA), presented using a computer running Windows Professional 7 (64 bit, 3 GB RAM, nVidia graphics card) and projected onto a screen placed 2.3 m from the participants (EPSON EMP-821 projector; refresh rate: 60 Hz, resolution: 1280 × 1024 pixels, projection screen size: 120 × 90 cm's). The stimuli and feedback symbols were subtended 4° × 4° and 3° × 3° of visual angle respectively.

Probabilistic reversal learning task. The experiment consisted of two blocks of 170 trials each, separated by a break. At the beginning of each block, participants were shown three symbols (A, B and C) chosen randomly from the larger set of twelve stimuli. The chosen symbols were used throughout the block. In each trial, participants were told to identify the symbol with the highest reward probability among a pair of stimuli selected from the three symbols. Each rewarded trial earned them 1 point, while unrewarded trials earned them zero points. Participants knew that their performance would be monitored and transformed into monetary rewards at the end of the experiment (up to a maximum of £45), without being instructed on the exact mapping between earned points and their final payoff.

At any given point during the experiment, one of the three symbols was associated with a 70% chance of obtaining a reward ("high" reward probability symbol) compared to the remaining two symbols, each of which had a 30% chance of obtaining a reward ("low" probability symbols). Participants were not informed of the exact reward probabilities assigned to the symbols and they were told instead to learn to choose the symbol that was more likely to lead to a reward through trial and error (i.e. making use of the outcome of past decisions). To prevent participants from searching for non-existent patterns and to reduce cognitive load we presented the three possible pair combinations of the three symbols in a fixed order (i.e. AB, BC and CA) – though the presentation on the screen (left or right of the fixation cross) for the two symbols was randomized. Participants were explicitly informed about this manipulation.

Each trial began with the presentation of a fixation cross for a random delay (1–4 s; mean 2.5 s). To minimize large eye-movements, participants were instructed to focus on the central fixation. Two of the three symbols were then placed to either side of the fixation cross for 1.25 s. During this period, participants had to press a button on a fORP MRI compatible response box (Current Design Inc., Philadelphia, PA, USA) using either their right index or middle finger to select the right or left symbol on the screen, respectively. The fixation cross flickered for 100 ms after each button press to signal to the participants that their response was registered. Finally, the decision outcome was presented after a second random delay (1–4 s; mean 2.5 s). Rewarded or unrewarded feedback was given by placing a tick or a cross, respectively, in the center of the screen for 0.65 s. Trials, in which participants failed to respond within the 1.25 s of the stimulus presentation, were followed by a "Lost trial" message and were excluded from further analyses. To increase estimation efficiency in the fMRI analysis, the timing of the two delay periods was optimized using a genetic algorithm[72,73]. Figure 1a summarizes the sequence of these events.

We defined a learning criterion for which participants were thought to have learned the task when they consistently selected the high probability symbol in 5 out of the last 6 trials. Every time this learning criterion was reached, we introduced a reversal by randomly re-assigning the "high" reward probability to a different symbol. This manipulation ensured that participants only experienced reversals after learning, as previously used in model-based fMRI studies[31,74]. To make reversals less predictable, we included additional buffer trials after the learning criterion was reached. The number of buffer trials followed a Poisson process, such that there was a probability of 0.3 that a reversal occurred on any subsequent trial (with a minimum of 1 and a maximum of 8 trials). Finally, a key component of this paradigm was that each stimulus pair was chosen from a larger set of three symbols. This manipulation encouraged participants to engage in an exploration phase to identify the most rewarding symbol after reversals and it forced the participants to choose between the two least rewarding symbols even after they had learned the task (when the two were presented together). This manipulation ensured a more balanced number of positive and negative RPEs across all trials.

Modeling of behavioral data. We used a model-free reinforcement learning (RL) algorithm to estimate trial-by-trial RPEs using each participants' behavioral choices[1]. Specifically, the algorithm assigned each choice i (for example selecting the symbol A) an expected value $v_A(i)$ which was updated via a RPE, $\delta(i)$, as follows:

$$v_A(i+1) = v_A(i) + \alpha.\ \delta(i) \tag{1}$$

where α is a fixed learning rate that determines the influence of the RPE on the updating of the stimulus expected value. The RPE was given by the following equation: $\delta(i) = r(i) - v_A(i)$, where $r\ (i)$ represents the observed outcome (0 or 1). The expected values of both the unselected stimulus (e.g. B) and the stimulus not shown on trial i (e.g. C) were not updated.

It is worth noting that subject-wise differences in overall task volatility (contingent upon the number of reversals attained during the task) were captured by different subject-wise estimates of the learning rate (for example, subjects who experienced a more stable environment – that is fewer reversals – had lower learning rate estimates). For comparison, to capture both subject- and trial-wise fluctuations in task volatility we also used a RL model incorporating a dynamic learning rate (DYNA) that enables a trial-by-trial scaling of the choice expected value updating[29]. In this model the learning rate (α) on each trial i is modulated by the slope of the smoothed unsigned RPE (m) according to the following update equations:

$$\begin{aligned} \alpha(i) &= \alpha(i-1) + f(m(i)).\ (1-\alpha(i-1)),\ \ if\ m > 0 \\ \alpha(i) &= \alpha(i-1) + f(m(i)).\ (\alpha(i-1)),\ \ if\ m < 0 \end{aligned} \tag{2}$$

where $f(m(i))$ is a double sigmoid function that transforms the slope m into the [0,1] interval and scales the trial-by-trial updating of the dynamic learning rate. Crucially, this transformation function is itself parameterized by a free parameter γ. High values of γ render the updating of the dynamic learning rate negligible so that in essence the learning rate becomes fixed. On inspection of the subject-wise fits of the dynamic learning rate model we found high values of the parameter γ across participants and unvarying trial-by-trial estimates of the dynamic learning rate ($\log(\gamma)$: mean = 2.87, SEM = 0.92). The Matlab code implementing the computational models described above is available at http://decision.ccni.gla.ac.uk/data/RLmodels.zip.

Whilst model-free RL approaches (such as the ones presented above) rest on the assumption that subjects make choices contingent upon the cached stimulus-expected value associations that have been acquired through prior experience, model-based RL approaches allow for representations of stimulus-outcome contingencies to bear on the decision process. To rule out that our subjects were merely inferring stimulus-outcome contingencies instead, we adapted the model presented in ref. 40 to our one stage task environment. Briefly, in our model-based RL model stimulus-outcome contingencies were updated as follows: $SO(i+1, s, r) = SO(i, s, r) + \alpha\delta(i)$ where SO represents a stimulus-outcome contingency matrix, s indicates the chosen stimulus, r specifies the type of binary outcome (1 for positive RPE and 0 for negative RPE), α is a fixed learning rate and δ is a stimulus-outcome prediction error computed as $\delta(i) = 1 - SO(i, s, r)$. The stimulus-outcome contingencies of the two stimuli that were not chosen and not shown on trial i were not updated.

As in the model-free variants of our RL model only the expected value of the chosen stimulus (e.g. A) was updated as $v_A(i+1) = SO(i+1, A,\ 1)$. In all models we used a softmax decision function in which, on each trial i, a stimulus choice probability (e.g. A) was given by:

$$P_A(i) = \sigma(\beta(v_A(i) - v_B(i)) - \phi) \tag{3}$$

where $\sigma(z) = 1/(1 + e^{-z})$ is the logistic function, ϕ denotes the indecision point (when choosing each of the alternatives is equiprobable) and β represents the degree of choice stochasticity (i.e., the exploration/exploitation parameter). Whilst choice probability of the unchosen stimulus (e.g B) was updated as follows: $P_B(i) = 1 - P_A(i)$, choice probability of the stimulus not shown on trial i (e.g. C) was not updated.

Model fitting and model comparison. For each subject j we estimated the set of model parameters θ_j using a *maximum likelihood estimation* fitting procedure: $\theta_j^{ML} = argmax_\theta\ logL\ \left(\theta_j\right)$. The log likelihood $logL$ was computed as $\frac{\sum C_a logP_a\left(\theta_j\right)}{N_a} + \frac{\sum C_b logP_b\left(\theta_j\right)}{N_b} + \frac{\sum C_c logP_c\left(\theta_j\right)}{N_c}$ where $log\ P\left(\theta_j\right)$ is the choice log-probability given the model parameters θ_j, C is a binary vector indexing observed choice, N is the number of observed choices and the subscripts indicate each of the 3 available choices. To preserve the parameters' natural bounds, log (β, γ) and logit (α) transforms of the parameters were implemented.

To determine the best fitting model we performed classical model comparison. Specifically, for each model we first estimated the subject-wise Bayesian Information Criterion (BIC) as follows: $BIC = -2log\ L + dlogn$. Here, the goodness of fit of a model ($-log\ L$) is penalised by the complexity term ($dlogn$) where the number of free parameters in the model d is scaled by the number of data points n (i.e. trials). We then computed the sum of subject-wise BIC for each model and compared the model-wise BIC estimates (lower estimates indicating better fit). We found the BIC for the model-free RL with a fixed learning rate to be the lowest ($BIC_{MF} = 413.26$; $BIC_{MB} = 413.61$; $BIC_{DYNA} = 529.17$).

To visualize the winning model goodness of fit we divided the subject-wise predicted choice probabilities into ten groups depending on the distribution quintiles and for each group computed the subject-wise average predicted choice probability. Using the sorting index of the predicted choice probabilities we then retrieved the subject- and group-wise average observed choice probabilities and computed Pearson's correlation coefficient (rho = 0.97, Fig. 1b).

Finally, we used the RPE estimates from the model-free RL with a fixed learning rate to provisionally separate trials into those with low-vs-high RPEs surprise to run a binary decoder on the EEG data (see below). Our main goal was to then exploit the single-trial variability in the electrophysiologically-derived (i.e. endogenous) and temporally-specific representations of RPE valence and surprise to build fMRI predictors in order to identify the networks associated with these representations in the human brain.

Electrophysiological recordings. EEG recordings were performed simultaneously with the fMRI acquisition. EEG data were acquired using Brain Vision Recorder (BVR; Version 1.10, Brain Products, Germany) and MR-compatible amplifiers (BrainAmps MR-Plus, Brain Products, Germany). The sampling rate was set at 5 kHz with a hardware bandpass filter of 0.016 to 250 Hz. The data were collected with an electrode cap consisting of 64 Ag/AgCl scalp electrodes (BrainCap MR; Brain Products, Germany) placed according to the international 10–20 system. Reference and ground electrodes were embedded within the EEG cap and positioned along the midline (reference: placed between electrodes Fpz and Fz, ground electrode: placed between electrode Pz and Oz). 10 kΩ surface-mount resistors were placed in line with each electrode for added subject safety, while all leads were twisted for their entire length and bundled together to minimize inductive pick-up. All input impedances were kept below 20 kΩ. Hardware-based EEG/MR clock synchronization was achieved using the SyncBox device (Syncbox, Brain Products, Germany) and MR-scanner triggers were collected to enable offline removal of MR gradient artifacts (GA). Scanner trigger pulses were stretched to 50 µs using an in-house pulse extender to facilitate accurate capture. These pulses along with all experimental event triggers were recorded by the BVR software to ensure synchronization with the EEG data.

To minimize the influence of the GA at source, electrodes Fp1 and Fp2 were positioned at the scanner's z = 0 position (i.e. the scanner's isocenter), corresponding to a 4 cm shift of the head in the foot-head direction[75]. This was achieved by aligning electrodes Fp1 and Fp2 with the laser beam used to position the subject inside the scanner. We used a 32-channel SENSE head coil with an access port that allowed all EEG cables to run along a straight path out of the MR, thereby ensuring no wire loops and minimizing the risk of RF heating the EEG cap. Finally we used a cantilevered beam to isolate the cables from scanner vibrations, thereby reducing induced artifacts as much as possible[76].

EEG pre-processing. We designed offline Matlab (Mathworks, Natick, MA) routines to perform basic processing of the EEG signals and remove GA and ballistocardiogram (BCG) artifacts (appearing due to magnetic induction on the EEG leads). The GA is inherently periodic and we therefore created an artifact template by averaging the EEG signal over 80 functional volumes and subsequently subtracting it from the EEG data centered on the middle volume. This procedure was repeated for each volume and each EEG channel separately. To remove any residual spike artifacts, we applied a 10 ms median filter. We then applied (non-causally to avoid distortions caused by phase delays) the following filters: a 0.5 Hz high-pass filter to remove DC drifts, 50 Hz and 100 Hz notch filters to remove electrical line noise, and 100 Hz low pass filter to remove high frequency signals not associated with neurophysiological processes of interest.

Finally, we treated the BCG artifacts, which often share frequency content with the EEG. In order to avoid loss of signal power that might otherwise be informative we adopted a conservative and previously validated approach[12,34]. Specifically, we only removed few subject-specific BCG components using principal component analysis and relied on our single-trial discriminant analysis (see below) to identify components that were likely to be orthogonal to the BCG (this is ultimately achieved due to the multivariate nature of our discriminant analysis). To extract the BCG principal components, we first low-pass filtered the EEG data at 4 Hz (e.g. the frequency range where BCG artifacts are observed), and then estimated subject-specific principal components. The average

number of components across subjects was 2.3. The sensor weightings corresponding to the BCG components were then projected onto the broadband data and subtracted out.

Eye-movement artifact removal. Before performing the main task, our participants were asked to complete an eye movement calibration task including blinking repeatedly when a fixation cross appeared in the center of the projection screen and making several horizontal and vertical saccades depending on the position of a fixation cross, subtending 0.6° × 0.6° of visual angle. Using principal component analysis we determined linear EEG sensor weightings corresponding to eye blinks and saccades such that these components were projected onto the broadband data from the main task and subtracted out[34,77].

EEG linear discriminant analysis. We used single-trial multivariate discriminant analysis of the EEG[12,32,78,79] to perform binary discriminations along the RPE valence and surprise dimensions. Specifically, for each participant, we estimated linear weightings of the EEG electrode signals (i.e. spatial filters) that maximally discriminated between (1) positive vs. negative feedback trials (valence dimension) and (2) high vs. low outcome surprise trials (surprise dimension) (Eq. 4). To identify temporally distinct neuronal components associated with each dimension we employed this procedure over several temporally distinct training windows. Applying the estimated spatial filters to single-trial data produced a measurement of the resultant discriminating component amplitudes. These values represent the distance of individual trials from the discriminating hyperplane and can be thought of as a surrogate for the neuronal response variability associated with each discriminating condition, with activity common to both conditions removed (Fig. 2a)[78].

Here, we defined discriminator training windows of interest with width 60 ms and center times τ, ranging from −100 to 600 ms relative to outcome onset (in 10 ms increments) and used a regularized Fisher discriminant to estimate a spatial weighting vector $w(\tau)$, which maximally discriminates between sensor array signals $x(t)$, for two conditions (i.e. positive vs negative RPEs and high vs low surprise RPE trials):

$$y_i(\tau) = \frac{1}{N}\sum_{t=\tau-N/2}^{t=\tau+N/2} w(\tau)^{\mathrm{T}} x_i(t) \tag{4}$$

where $x(t)$ is an D x T matrix (D sensors and T time samples) and $y_i(\tau)$ represents the resulting single-trial discriminator amplitudes. In separating the relevant groups of trials for RPE valence and surprise the discriminator was designed to map positive RPE and high surprise RPE trials to positive amplitudes and negative RPE and low surprise RPE trials to negative amplitudes.

Our aim was to capitalize on the endogenous variability in these single-trial discriminator amplitudes ($y_i(\tau)$) to build EEG-informed BOLD predictors (Fig. 2b) for analyzing the simultaneously acquired fMRI data (see below). Our working hypothesis is that single-trial variability in our electrophysiologically-derived measures of RPE valence and surprise can enable otherwise "static" fMRI activations (resulting from temporal averaging and the slow dynamics of conventional fMRI) to be absorbed by temporally specific components, thereby offering a more complete spatiotemporal picture of the underlying networks. Note that the trial-by-trial variability in our EEG component amplitudes is driven mostly by cortical regions in close proximity to our recording sensors and to a lesser extend by distant (e.g. subcortical) structures. Nonetheless, as long as the BOLD activity of relevant subcortical regions covaries together with the BOLD signal in the cortical sources of our EEG activity, then the same EEG-informed fMRI predictors that revealed the cortical structures will also absorb the deeper structures (regardless of whether the latter contributed directly to the EEG signal)[80–82].

We estimated the spatial vectors $w(\tau)$ in Equation 4 for each time window τ as follows: $w = S_c(m_2 - m_1)$ where m_i is the estimated mean for condition i and $S_c = 1/2(S_1 + S_2)$ is the estimated common covariance matrix (i.e. the average of the empirical covariance matrices for each condition, $S_i = 1/(N-1)\sum_{j=1}^{N}(x_j - m_i)(x_j - m_i)^T$, with N = number of trials). To treat potential estimation errors we replaced the condition-wise covariance matrices with regularized versions: $\widetilde{S}_i = (1-\lambda)S_i + \lambda\nu I$, with $\lambda \in [0, 1]$ being the regularization term and ν the average eigenvalue of the original S_i (i.e. trace $(S_i)/D$). Note that $\lambda = 0$ yields unregularized estimation and $\lambda = 1$ assumes spherical covariance matrices. We optimized λ for each subject separately during the entire period following the outcome, using a leave-one-out trial cross validation (λ's, mean ± se: 0.028 ± 0.05).

To quantify the discriminator's performance at each time window, we calculated the area under a receiver operating characteristic curve, also known as Az value, using a leave-one-out trial cross validation[83]. Next, we assessed the significance of the discrimination performance using a permutation test by performing the leave-one-out trial procedure after randomizing the labels associated with each trial. To produce a probability distribution for Az, and estimate the Az value leading to a significance level of $P < 0.01$, we repeated this randomization procedure 1000 times. Note that our classification results were virtually identical when using a logistic regression approach to train the classifier[33].

Given the linearity of our approach, we also computed scalp topographies for each discriminating component of interest resulting from Equation 4 by estimating a forward model as:

$$a(\tau) = \frac{x(\tau)y(\tau)}{y(\tau)^T y(\tau)} \tag{5}$$

where $y_i(\tau)$ is now shown as a vector $y(\tau)$, where each row being a trial, and $x_i(t)$ is organized as a matrix, $x(\tau)$, with rows as channels and columns as trials, all for time window τ. These forward models can be viewed as scalp

plots and interpreted as the coupling between the discriminating components and the observed EEG[12,33,34,79]. The code for the multivariate discriminant analysis is available at http://sccn.ucsd.edu/eeglab/plugins/lr1.2.zip.

Mixed-effects analyses of behavioral and EEG data. To account for between subjects variability, we fitted mixed-effects regression models including by-subject random intercept and by-subject random slopes for all predictors of interest using the lme4 package (https://cran.r-project.org/web/ packages/lme4/index.html) in R (http://www.r-project.org). Details of the dependent and predictor variables used for each regression analysis are provided in the main text. To assess significance of the fixed-effects we performed an F-test using the Satterwaite/Kenward-Roger approximation of denominator's degrees of freedom (linear regression) and a likelihood ratio test (logistic regression). Finally, the significance of a predictor variable or set of variables is tested using a log-likelihood ratio test, whereby the log-likelihood of the model with all predictors is compared with the log-likelihood of the model without the predictors being tested. The difference in the log-likelihood of two models is distributed according to a χ^2 distribution whose degrees of freedom equal the difference in the number of parameters in the two models.

MRI data acquisition. MRI data were acquired on a 3 Tesla Philips Achieva MRI scanner (Philips, Netherlands). Our functional MRI data were collected using a 32-channel SENSE head coil (SENSE factor = 2.3) measuring 40 slices of 80 × 80 voxels (3 mm isotropic), a field of view (FOV) of 204 mm, flip angle (FA) of 80°, repetition time (TR) of 2.5 s and echo time (TE) of 40 ms. Slices were acquired in interleaved fashion. In total, two runs of 468 functional volumes, corresponding to the blocks of trials in the main task, were acquired. To correct for B0 inhomogeneities in the fMRI data, a B0 map was also acquired using a multi-shot gradient echo sequence (32 slices of 80 × 80 voxels – 3 mm isotropic, FOV: 204 mm, FA: 90°, TR: 383 ms, TE: 2.3 ms, delta TE: 5 ms). Anatomical images were collected using a MPRAGE T1-weighted sequence (160 slices of 256 × 256 voxels – 1 mm isotropic, FOV: 256 mm, TR: 8.2 ms, TE: 3.7 ms).

fMRI data preprocessing. The first five volumes from each experimental run were removed to ensure that steady-state imaging has been achieved. We used the remaining 463 volumes for all subsequent statistical analyses. fMRI data preprocessing included motion correction, slice-timing correction, high-pass filtering (> 100 s) and spatial smoothing (with a Gaussian kernel of 8 mm full-width at half maximum). These steps were achieved using the FMRIB's Software Library (Functional MRI of the Brain, Oxford, UK). Next, we applied B0 unwarping onto the fMRI images to correct for signal loss and geometric distortions due to B0 field inhomogeneities. Finally, the registration of fMRI data to standard space (Montreal Neurological Institute, MNI) was performed using FMRIB's Non-linear Image Registration Tool using a 10 mm warp resolution. This procedure included an initial linear transformation of the fMRI images into an individual's high-resolution space (using six degrees of freedom) prior to applying the non-linear transformation to standard space.

fMRI analysis. We employed a multilevel approach within the general linear model (GLM) framework to perform whole-brain statistical analyses of functional data as implemented in FSL[84]:

$$Y = X\beta + \varepsilon = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N + \varepsilon \quad (6)$$

where Y is a $T \times 1$ (T time samples) column vector containing the times series data for a given voxel, and X is a $T \times N$ (N regressors) design matrix with columns representing each of the psychological regressors convolved with a canonical hemodynamic response function (double-γ function). β is a $N \times 1$ column vector of regression coefficients and ε a $T \times 1$ column vector of residual error terms. Using this model we initially performed a first-level fixed effects analysis to process each participant's individual experimental runs, which were then combined in a second-level fixed effects analysis. Finally, we combined results across participants in a third-level, mixed-effects model (FLAME 1), treating subjects as a random effect. Time series statistical analysis was carried out using FMRIB's improved linear model with local autocorrelation correction. Applying this framework, we performed the GLMs highlighted below.

GLM1 – EEG-informed fMRI analysis of outcome phase. Our main fMRI analysis was designed to reveal the brain networks underlying the two main outcome dimensions – valence and surprise – by capitalizing on the endogenous STV in temporally-specific EEG components associated with these dimensions. For the valence dimension, we aimed at exposing brain areas in which the BOLD signal varied with the EEG STV along a positive versus negative axis, while for the surprise dimension the BOLD signal varied with the EEG STV along a high versus low surprise RPE axis, regardless of valence.

Specifically, locked to the time of outcome (i.e. when the tick/cross appeared) we included five boxcar regressor with a duration of 100 ms for each regressor event: 1) an unmodulated regressor (UM; all event amplitudes set to one), 2) a fully parametric regressor whose event amplitudes were modulated by the EEG STV associated with the RPE valence component of interest ($EEG_{LateVal}$), 3) a fully parametric regressor whose event amplitudes were modulated by the outcome surprise estimates from the RL model ($MODEL_{Sur}$), and 4) a fully parametric regressor whose event amplitudes were modulated by the EEG STV associated with the RPE surprise component of interest ($EEG_{LateSur}$). Finally, we included a fully parametric regressor whose event amplitudes were modulated by the EEG STV associated with an earlier RPE valence component (to account for fast alertness response to negative RPEs as per[12], $EEG_{EarlyVal}$), an unmodulated regressor for all lost trials at time of outcome ($LOST$), an unmodulated regressor of no interest at the time of stimulus presentation (i.e. decision phase, DEC), and six nuisance regressors, one for each of the motion parameters (three rotations and three translations), such as:

$$Y = \beta_1 UM + \beta_2 EEG_{LateVal} + \beta_3 MODEL_{Sur} + \beta_4 EEG_{LateSur} + \beta_5 EEG_{EarlyVal} + \beta_6 LOST + \beta_7 DEC \quad (7)$$

Note that the amplitudes of the two valence regressors ($EEG_{LateVal}$ and $EEG_{EarlyVal}$) were largely uncorrelated[12].

To examine the extent to which neural representations of RPE valence and surprise overlap spatially across participants, we performed a conjunction analysis on the results arising from the EEG-informed regressors for the late valence and surprise components. Specifically, we examined which brain areas were jointly activated by creating an intersection of statistical maps relating to the two components (as implemented in *easythresh_conj* within FSL[85]). The conjunction analysis was performed at the whole-brain level and the resulting statistical image was thresholded at Z-score > 2.57 with a cluster probability threshold of P = 0.05.

GLM2 – Region of interest analysis. To visualize the overall response profile of regions of interest (ROIs) (e.g. those representing both valence and surprise from GLM1) we ran another model with trials binned into six outcome-locked regressors, separated into positive and negative RPEs as well as according to three surprise groups for each outcome type (low, medium, high) using the EEG STV in our outcome surprise component (0–33%, 33–66%, and 66–100% percentiles of EEG amplitudes). The main motivation for this analysis was to visualize the response profile within these areas with respect to the strength of the endogenous variability carried by the EEG STV (rather than the variability from unsigned RPE amplitudes from the RL model). In addition we included the same regressors of no interest we used in GLM1 above.

From each of the six regressors we extracted beta coefficients from the following ROIs: ventromedial prefrontal cortex (vmPFC), striatum (STR), middle temporal gyrus (MTG) and lingual gyrus (LG). More specifically, we first extracted ROIs masks at the group level from GLM1 (i.e. in standard space), by applying the cluster correction procedure described above. We subsequently back-projected these ROIs from standard space into each individual's EPI (functional) space by applying the inverse transformations as estimated during registration (see *fMRI data preprocessing* section). Each ROI was then checked against the relevant (regressor-specific) statistical maps in the individual brains [at a slightly more lenient threshold of P < 0.01 uncorrected, cluster size > 10 voxels (90 mm^3)] to ensure that the inverse-transformation was performed properly. Finally, for each of the six regressors we computed average beta coefficients from all voxels in the back-projected clusters and across participants to visualize the overall response profile of the ROIs as a function of both RPE valence and surprise.

Relationship between percent signal change and value updating. To establish a more concrete link between the brain regions encoding both outcome-related signals, we ran an additional regression analysis. Specifically, we directly tested the relationship between trial-related percent signal change (PSC) in those areas and stimulus value updating, the key underlying component of learning, defined as the difference in successive prediction values, $v^k(i+1) - v^k(i)$, computed using estimates of our RL model, with k indexing a given choice and with i indexing trial. PSC was estimated on outcome-locked data by defining a temporal window extending 5 s (2 volumes) before the outcome and extending to 10 s (4 volumes) after the outcome onset. We then estimated the PSC traces for all outcomes as follows:

$$PSC_i(t) = 100 \times (X_i(t) - X_i^b)/\overline{X} \quad (8)$$

where $X(t)$ is the mean outcome-locked data at time point t, X^b is the mean baseline signal defined as the average signal during the 5 s preceding each outcome and extending two volumes after onset, and $\overline{X}$ the mean outcome-locked volume signal across all data points in a given run. Finally, we employed the PSC as our predictor of value update in a linear regression such as:

$$value\_update = \beta_0 + \beta_1 \times PSC + \varepsilon \quad (9)$$

Then, to establish a significant trial-by-trial association between PSC in our ROIs and value update, we tested whether the regression coefficients resulting from all subjects (β_1's) come from a distribution with mean greater than zero (using a one-tailed *t-test*).

Resampling procedure for fMRI thresholding. In line with recent recommendations and critical developments in the literature[86], we corrected the statistical maps for multiple comparisons by employing a resampling procedure that considers the a priori statistics of the single-trial variability in all of our parametric predictors in a way that trades off cluster size and maximum voxel Z-score[12]. Specifically, we preserved the overall distributions of the EEG discriminating components ($y_i(\tau)$ for the RPE valence and surprise components) as well as the single-trial variability of the model-based unsigned RPE predictor while removing the specific single-trial correlations in subject-specific experimental runs. Thus, for each resampled iteration and each predictor, all trials were built upon the original regressor amplitude distribution; however the specific values were mixed across trials and runs. In other words, the same regressor amplitudes were used in each permutation, however, the sequence of regressor events was randomized.

We repeated this resampling procedure 100 times for each participant, including a full three-level analysis (run, subject, and group). The design matrix included the same regressors of non-interest used in all our GLM analyses. Consequently, this process enabled us to construct the null hypothesis H0, and establish a joint threshold on cluster size and Z-score based on the cluster outputs from the randomized predictors. We then extracted cluster sizes from all activations with |Z|-score > 2.57 for both positive and negative correlations with the permuted predictors. Finally, we produced a distribution for the cluster sizes for the permuted data, and estimated the largest cluster size leading to a significance level of P < 0.05. This procedure resulted in a corrected threshold for our statistical maps, which we then applied to the activations observed in the original data. This cluster-threshold was set to > 76 voxels at $|Z| = 2.57$.

References

1. Sutton, R. Reinforcement Learning: *An Introduction*. (MIT Press, 1998).
2. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275**, 1593–1599 (1997).
3. Fiorillo, C. D., Tobler, P. N. & Schultz, W. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898–1902 (2003).
4. Chau, B. K. H. *et al*. Contrasting Roles for Orbitofrontal Cortex and Amygdala in Credit Assignment and Learning in Macaques. *Neuron* **87**, 1106–1118 (2015).
5. Niv, Y. *et al*. Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *J. Neurosci.* **35**, 8145–8157 (2015).
6. Delgado, M. R., Miller, M. M., Inati, S. & Phelps, E. A. An fMRI study of reward-related probability learning. *NeuroImage* **24**, 862–873 (2005).
7. Iglesias, S. *et al*. Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron* **80**, 519–530 (2013).
8. Bach, D. R., Hulme, O., Penny, W. D. & Dolan, R. J. The Known Unknowns: Neural Representation of Second-Order Uncertainty, and Ambiguity. *J. Neurosci.* **31**, 4811–4820 (2011).
9. Ide, J. S., Shenoy, P., Yu, A. J. & Li, C. R. Bayesian Prediction and Evaluation in the Anterior Cingulate Cortex. *J. Neurosci. Off. J. Soc. Neurosci.* **33**, 2039–2047 (2013).
10. Frank, M. J. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *J. Cogn. Neurosci.* **17**, 51–72 (2005).
11. Wunderlich, K., Rangel, A. & O'Doherty, J. P. Neural computations underlying action-based decision making in the human brain. *Proc. Natl. Acad. Sci.* **106**, 17199–17204 (2009).
12. Fouragnan, E., Retzler, C., Mullinger, K. & Philiastides, M. G. Two spatiotemporally distinct value systems shape reward-based learning in the human brain. *Nat. Commun.* **6**, (2015).
13. Dayan, P., Kakade, S. & Montague, P. R. Learning and selective attention. *Nat Neurosci* **3**, 1218–1223 (2000).
14. Pearce, J. M. & Hall, G. A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980).
15. Kruschke, J. K. Toward a Unified Model of Attention in Associative Learning. *J. Math. Psychol.* **45**, 812–863 (2001).
16. Roesch, M. R., Esber, G. R., Li, J., Daw, N. D. & Schoenbaum, G. Surprise! Neural Correlates of Pearce-Hall and Rescorla-Wagner Coexist within the Brain. *Eur. J. Neurosci* **35**, 1190–1200 (2012).
17. Pearce, J. M. & Mackintosh, N. J. in *A*ttention and *A*ssoc*iative Learning: From Brain to Behaviour* (eds Mitchell, C. & Le Pelley, M. E.) 11–39 (Oxford University Press, 2010).
18. den Ouden, H. E. M., Kok, P. & de Lange, F. P. How Prediction Errors Shape Perception, Attention, and Motivation. *Front. Psychol.* **3** (2012).
19. Jensen, J. *et al*. Separate brain regions code for salience vs. valence during reward prediction in humans. *Hum. Brain Mapp.* **28**, 294–302 (2007).
20. Knutson, B., Adams, C. M., Fong, G. W. & Hommer, D. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci. Off. J. Soc. Neurosci.* **21**, RC159 (2001).
21. Matsumoto, M. & Hikosaka, O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**, 837–841 (2009).
22. Fiorillo, C. D. Two Dimensions of Value: Dopamine Neurons Represent Reward But Not Aversiveness. *Science* **341**, 546–549 (2013).
23. Zink, C. F., Pagnoni, G., Martin, M. E., Dhamala, M. & Berns, G. S. Human striatal response to salient nonrewarding stimuli. *J. Neurosci. Off. J. Soc. Neurosci.* **23**, 8092–8097 (2003).
24. den Ouden, H. E. M., Daunizeau, J., Roiser, J., Friston, K. J. & Stephan, K. E. Striatal prediction error modulates cortical coupling. *J. Neurosci. Off. J. Soc. Neurosci.* **30**, 3210–3219 (2010).
25. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
26. Preuschoff, K., Bossaerts, P. & Quartz, S. R. Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* **51**, 381–390 (2006).
27. Metereau, E. & Dreher, J.-C. Cerebral correlates of salient prediction error for different rewards and punishments. *Cereb. Cortex N. Y. N 1991* **23**, 477–487 (2013).
28. Collins, A. G. E. & Frank, M. J. Surprise! Dopamine signals mix action, value and error. *Nat. Neurosci.* **19**, 3–5 (2016).
29. Rouhani, N., Norman K. A. & Niv, Y. Dissociable effects of surprising rewards on learning and memory. *BioRxiv Prepr*. doi:http://dx.doi.org/10.1101/111070
30. Philiastides, M. G., Biele, G., Vavatzanidis, N., Kazzer, P. & Heekeren, H. R. Temporal dynamics of prediction error processing during reward-based decision making. *NeuroImage* **53**, 221–232 (2010).
31. O'Doherty, J., Critchley, H., Deichmann, R. & Dolan, R. J. Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *J. Neurosci. Off. J. Soc. Neurosci.* **23**, 7931–7939 (2003).
32. Philiastides, M. G. & Sajda, P. EEG-Informed fMRI Reveals Spatiotemporal Characteristics of Perceptual Decision Making. *J. Neurosci.* **27**, 13082–13091 (2007).
33. Parra, L. C., Spence, C. D., Gerson, A. D. & Sajda, P. Recipes for the linear analysis of EEG. *NeuroImage* **28**, 326–341 (2005).
34. Goldman, R. I. *et al*. Single-trial discrimination for integrating simultaneous EEG and fMRI: Identifying cortical areas contributing to trial-to-trial variability in the auditory oddball task. *NeuroImage* **47**, 136–147 (2009).
35. Krugel, L. K., Biele, G., Mohr, P. N. C., Li, S.-C. & Heekeren, H. R. Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proc. Natl. Acad. Sci* **106**, 17951–17956 (2009).
36. Dayan, P. & Niv, Y. Reinforcement learning: The Good, The Bad and The Ugly. *Curr. Opin. Neurobiol.* **18**, 185–196 (2008).
37. Hayden, B. Y., Heilbronner, S. R., Pearson, J. M. & Platt, M. L. Surprise Signals in Anterior Cingulate Cortex: Neuronal Encoding of Unsigned Reward Prediction Errors Driving Adjustment in Behavior. *J. Neurosci.* **31**, 4178–4187 (2011).
38. Philiastides, M. G., Biele, G. & Heekeren, H. R. A mechanistic account of value computation in the human brain. *Proc. Natl. Acad. Sci.* **107**, 9430–9435 (2010).
39. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
40. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
41. Palminteri, S., Khamassi, M., Joffily, M. & Coricelli, G. Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* **6** (2015).
42. Bossaerts, P. Risk and risk prediction error signals in anterior insula. *Brain Struct. Funct.* **214**, 645–653 (2010).
43. Rudorf, S., Preuschoff, K. & Weber, B. Neural Correlates of Anticipation Risk Reflect Risk Preferences. *J. Neurosci.* **32**, 16683–16692 (2012).
44. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci. Off. J. Soc. Neurosci.* **16**, 1936–1947 (1996).
45. Yu, A. J. & Dayan, P. Uncertainty, Neuromodulation, and Attention. *Neuron* **46**, 681–692 (2005).

46. Ruff, C. C. & Fehr, E. The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurosci.* **15**, 549–562 (2014).
47. Kahnt, T. & Tobler, P. N. Salience Signals in the Right Temporoparietal Junction Facilitate Value-Based Decisions. *J. Neurosci.* **33**, 863–869 (2013).
48. Asaad, W. F. & Eskandar, E. N. Encoding of Both Positive and Negative Reward Prediction Errors by Neurons of the Primate Lateral Prefrontal Cortex and Caudate Nucleus. *J. Neurosci.* **31**, 17772–17787 (2011).
49. Sambrook, T. D. & Goslin, J. Mediofrontal event-related potentials in response to positive, negative and unsigned prediction errors. *Neuropsychologia* **61**, 1–10 (2014).
50. Seymour, B., Daw, N., Dayan, P., Singer, T. & Dolan, R. Differential Encoding of Losses and Gains in the Human Striatum. *J. Neurosci.* **27**, 4826–4831 (2007).
51. Foerde, K., Race, E., Verfaellie, M. & Shohamy, D. A Role for the Medial Temporal Lobe in Feedback-Driven Learning: Evidence from Amnesia. *J. Neurosci.* **33**, 5698–5704 (2013).
52. Litt, A., Plassmann, H., Shiv, B. & Rangel, A. Dissociating Valuation and Saliency Signals during Decision-Making. *Cereb. Cortex* **21**, 95–102 (2011).
53. Rushworth, M. F. S., Noonan, M. P., Boorman, E. D., Walton, M. E. & Behrens, T. E. Frontal Cortex and Reward-Guided Learning and Decision-Making. *Neuron* **70**, 1054–1069 (2011).
54. Dorris, M. C. & Glimcher, P. W. Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* **44**, 365–378 (2004).
55. Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Choosing the greater of two goods: neural currencies for valuation and decision making. *Nat. Rev. Neurosci.* **6**, 363–375 (2005).
56. Lim, S.-L., O'Doherty, J. P. & Rangel, A. The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. *J. Neurosci. Off. J. Soc. Neurosci.* **31**, 13214–13223 (2011).
57. Kim, H. & Cabeza, R. Trusting our memories: dissociating the neural correlates of confidence in veridical versus illusory memories. *J. Neurosci. Off. J. Soc. Neurosci.* **27**, 12190–12197 (2007).
58. McGuire, J. T., Nassar, M. R., Gold, J. I. & Kable, J. W. Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron* **84**, 870–881 (2014).
59. Akaishi, R., Kolling, N., Brown, J. W. & Rushworth, M. Neural Mechanisms of Credit Assignment in a Multicue Environment. *J. Neurosci.* **36**, 1096–1112 (2016).
60. Gläscher, J., Hampton, A. N. & O'Doherty, J. P. Determining a Role for Ventromedial Prefrontal Cortex in Encoding Action-Based Value Signals During Reward-Related Decision Making. *Cereb. Cortex* **19**, 483–495 (2009).
61. Kolling, N., Wittmann, M. & Rushworth, M. F. S. Multiple Neural Mechanisms of Decision Making and Their Competition under Changing Risk Pressure. *Neuron* **81**, 1190–1202 (2014).
62. Schultz, W. Dopamine reward prediction-error signalling: a two-component response. *Nat. Rev. Neurosci.* **17**, 183–195 (2016).
63. Ikemoto, S. Dopamine reward circuitry: Two projection systems from the ventral midbrain to the nucleus accumbens–olfactory tubercle complex. *Brain Res. Rev.* **56**, 27–78 (2007).
64. Matsumoto, M., Matsumoto, K., Abe, H. & Tanaka, K. Medial prefrontal cell activity signaling prediction errors of action values. *Nat. Neurosci.* **10**, 647–656 (2007).
65. Atallah, H. E., McCool, A. D., Howe, M. W. & Graybiel, A. M. Neurons in the Ventral Striatum Exhibit Cell-Type-Specific Representations of Outcome during Learning. *Neuron* **82**, 1145–1156 (2014).
66. Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C. & Fiez, J. A. Tracking the Hemodynamic Responses to Reward and Punishment in the Striatum. *J. Neurophysiol.* **84**, 3072–3077 (2000).
67. D'Ardenne, K., McClure, S. M., Nystrom, L. E. & Cohen, J. D. BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science* **319**, 1264–1267 (2008).
68. Caplin, A., Dean, M., Glimcher, P. W. & Rutledge, R. B. Measuring Beliefs And Rewards: A Neuroeconomic Approach. *Q. J. Econ.* **125**, 923–960 (2010).
69. Rutledge, R. B., Dean, M., Caplin, A. & Glimcher, P. W. Testing the Reward Prediction Error Hypothesis with an Axiomatic Model. *J. Neurosci. Off. J. Soc. Neurosci.* **30**, 13525–13536 (2010).
70. Stenner, M.-P. *et al*. No unified reward prediction error in local field potentials from the human nucleus accumbens: evidence from epilepsy patients. *J. Neurophysiol.* **114**, 781–792 (2015).
71. Kishida, K. T. *et al*. Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proc. Natl. Acad. Sci.* **113**, 200–205 (2016).
72. Friston, K. J., Zarahn, E., Josephs, O., Henson, R. N. & Dale, A. M. Stochastic designs in event-related fMRI. *NeuroImage* **10**, 607–619 (1999).
73. Wager, T. D. & Nichols, T. E. *Optimization of experimental design in fMRI: a general framework using a genetic algorithm*. (2003).
74. O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J. & Andrews, C. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat. Neurosci.* **4**, 95–102 (2001).
75. Mullinger, K. J., Yan, W. X. & Bowtell, R. Reducing the gradient artefact in simultaneous EEG-fMRI by adjusting the subject's axial position. *NeuroImage* **54**, 1942–1950 (2011).
76. Mullinger, K. J., Castellone, P. & Bowtell, R. Best Current Practice for Obtaining High Quality EEG Data During Simultaneous fMRI. *J. Vis. Exp. JoVE*. doi:10.3791/50283 (2013).
77. Gherman, S. & Philiastides, M. G. Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage* **106**, 134–143 (2015).
78. Philiastides, M. G. & Sajda, P. Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cereb. Cortex N. Y. N 1991* **16**, 509–518 (2006).
79. Philiastides, M. G., Heekeren, H. R. & Sajda, P. Human Scalp Potentials Reflect a Mixture of Decision-Related Signals during Perceptual Choices. *J. Neurosci.* **34**, 16877–16889 (2014).
80. Plichta, M. M. *et al*. Simultaneous EEG and fMRI reveals a causally connected subcortical-cortical network during reward anticipation. *J. Neurosci. Off. J. Soc. Neurosci.* **33**, 14526–14533 (2013).
81. Walz, J. M. *et al*. Simultaneous EEG-fMRI reveals temporal evolution of coupling between supramodal cortical attention networks and the brainstem. *J. Neurosci. Off. J. Soc. Neurosci.* **33**, 19212–19222 (2013).
82. Keynan, J. N. *et al*. Limbic Activity Modulation Guided by Functional Magnetic Resonance Imaging-Inspired Electroencephalography Improves Implicit Emotion Regulation. *Biol. Psychiatry* **80**, 490–496 (2016).
83. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification*. (Wiley-Interscience, 2000).
84. Smith, S. M. *et al*. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23**(Suppl 1), S208–219 (2004).
85. Nichols, T., Brett, M., Andersson, J., Wager, T. & Poline, J.-B. Valid conjunction inference with the minimum statistic. *NeuroImage* **25**, 653–660 (2005).
86. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci.* **113**, 7900–7905 (2016).

Acknowledgements
This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC; grants BB/J015393/1–2 to MGP).

Author Contributions
E.F. and M.G.P. designed the experiments. E.F., C.R. and K.J.M. performed the experiments. E.F., F.Q and M.G.P. analyzed the data. E.F., F.Q. and M.G.P. wrote the paper. All authors discussed the results and implications and commented on the manuscript at all stages.

Additional Information
Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.